Uniwersytet Jagielloński
Wydział Matematyki i Informatyki
Instytut Matematyki

Person, Michał Bujak

Credit risk

HSBC Quant Academy
Project no. 2

Kraków 2021

Contents

1 Introduction **2**

2 Data analysis **3**

2.1 Variable selection 3

2.2 Data splitting 5

3 Logistic regression **6**

3.1 Full model 6

3.2 Analyst's model 7

3.3 Automatised model 7

3.4 Cross-Validation 9

3.5 Probit regression 9

4 Linear regression **10**

4.1 Automatised model 11

4.2 Analyst's model 11

4.3 Experts' model 12

5 Model comparison **12**

5.1 ROC and AUC 12

5.2 Model comparison 13

6 References **15**

7 Appendix A **16**

8 Appendix B **20**

1 Introduction

The main purpose of this project is to build and apply various types of prediction models that assess the probability of default for certain Middle-Market Wholesale companies. Initial stage consists of data quality assessment and univariate and multivariate analysis. Significant part of research is based on building logistic, probit and linear regression models, which is followed by profound validation of these objects. Comparison of effectiveness and performance is eventually conducted in order to choose the most valid model from the statistical point of view. The whole process is based on data corresponding to a set of qualitative and quantitative information collected from Credit Risk team in the period $2000 - 2008$ for Middle-Market Wholesale customers.
Data

- CUSTOMER_ID - internal customer identification number.
- ASSESSMENT_DEMAND - year of the credit expert assessment
- PRODUCT_DEMAND - credit expert's opinion on the competitive environment where the company operates, including its market position and the quality of its portfolio. The variable can show values from 10 to 90, 90 being the best score a company may obtain.
- OWNERS_MANAGEMENT - credit expert's opinion on the quality of the management of the company. The variable can show values from 10 to 90, 90 being the best score a company may obtain.
- ACCESS_CREDIT - credit expert's opinion on the ability of the company to generate profits based on its current portfolio. The variable can show values from 10 to 90, 90 being the best score a company may obtain.
- PROFITABILITY - credit expert's opinion on the ability of the company to generate profits based on its current portfolio. The variable can show values from 10 to 90, 90 being the best score a company may obtain.
- SHORT_TERM_LIQUIDITY - credit expert's opinion on the ability of the company to generate cash-flows in the short-term to fulfill short-term financial obligations. The variable can show values from 10 to 90, 90 being the best score a company may obtain.
- MEDIUM_TERM_LIQUIDITY - credit expert's opinion on the ability of the company to generate cash-flows in the medium and long-term. The variable can show values from 10 to 90, 90 being the best score a company may obtain.
- GROUP_FLAG - categorical variable which indicates whether the customer belongs to a Financial Holding. It may have two values: ”0” - The counterparty does not belong to a Financial Holding,”1” - The counterparty belongs to a Financial Holding.
- TURNOVER - it contains the value of the Turnover reported in the financial statements available for the assessment.
- INDUSTRY - it contains the industry in which the company operates.

- DEFAULT_FLAG - it contains whether the customer has gone into default in a 12-month period after the credit expert's assessment.

2 Data analysis

2.1 Variable selection

First thing we have to check is the variable CUSTOMER_ID to assess if we deal with repeated values.

By referring to table 1 we can clearly see that some of the companies applied for credit expert assessment more than once. However, all the observations are relevant, because factors reasoning (dis)approval may change and consequently clients are somehow independent of themselves with respect to time. At that point, there are two variables that we are not going to use in the process of building models - previously considered CUSTOMER_ID and ASSESSMENT_DEMAND, which are irrelevant for our future analysis from substantive point of view. CUSTOMER_ID is just an index variable, while ASSESSMENT_DEMAND does not matter, as the regressand DEFAULT_FLAG is independent from the year of the credit expert assessment.

Number of ID repetitions	Number of companies
1	3674
2	782
3	139
4	27
5	5
6	1
7	1

Table 1: Repetitions in CUSTOMER_ID variable

Now we should analyse the distribution of our continuous variables with respect to the dependant variable.

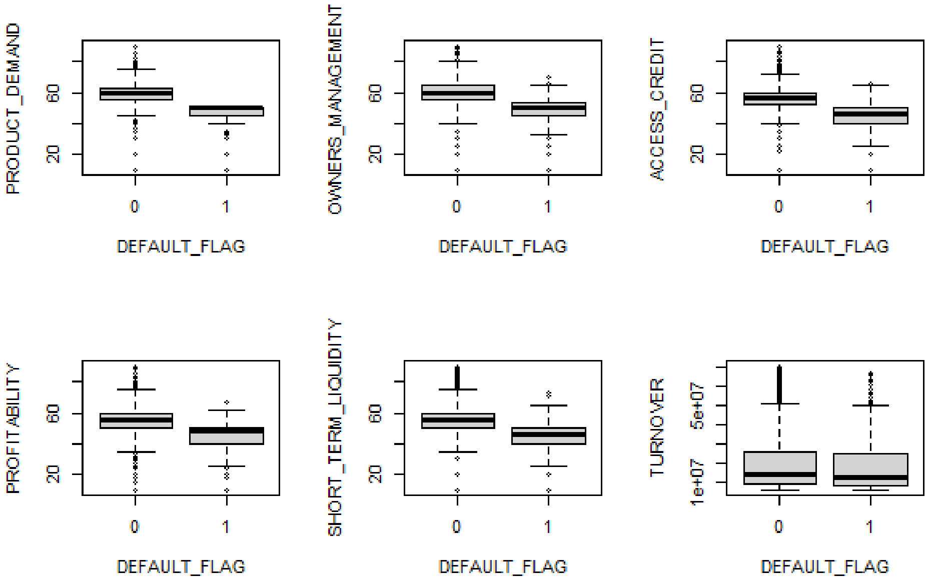


Figure 1: Distribution of continuous regressors with respect to regressand

It is clearly visible that most of the variables from image 1 are valuable from the informative perspective, which means that their distributions give us information about levels of DEFAULT_FLAG. The variable TURNOVER may be unserviceable as its values are practically identical with respect to the levels of regressand, but at this point we decide not to remove it yet.

Next useful statistic tool for checking the informative power of our independant variables is Information Value. This statistic is used when working with binary regressand like ours. To evaluate its value for each regressor, first we have to divide it into bins, if the variable is continuous or just simply use levels of factor variable as a bins. Next step is to calculate the positive and negative distribution for each bin, by using the corresponding values of our dependant variable. If we denote them respectively as DP_i and DN_i (where both are fractions with nominators as number of positive/negative corresponding values in bin i and denominators as number of all positive/negative values), then the Information Value formula is as in formula 1:

$$IV = \sum_i (DN_i - DP_i) ln(\frac{DN_i}{DP_i}). \tag{1}$$

The logarithm component in formula 1 is known as Weight of Evidence.

Variable	Information Value
PRODUCT_DEMAND	5.2142808988
ACCESS_CREDIT	2.7631735048
OWNERS_MANAGEMENT	2.4141578359
SHORT_TERM_LIQUIDITY	2.3368549332
MEDIUM_TERM_LIQUIDITY	2.1336758455
PROFITABILITY	1.9460299107
INDUSTRY	0.2136398385
TURNOVER	0.0733426738
GROUP_FLAG	0.0002749331

Table 2: Information Value of regressors

The interpretation of this value is straightforward - values greater than 0.3 are considered to be strong predictors, those from between 0.1 and 0.3 are medium predictors, those from below 0.1 are weak predictors and finally variables with IV lower than 0.02 are unserviceable. Therefore judging by the values from table 2, we can see that variable GROUP_FLAG is not suitable for regressor, while TURNOVER is a weak predictor. Taking under consideration that TURNOVER variable presented badly during visual analysis of distribution, we decide to discard both variables from our future model building. Detailed graphics concerning IV are presented in Appendix A (section 7).

An important part of the variable selection process is checking whether they have linear correlations between each other. Strong linear connections between regressors are a highly unwanted phenomenon, which is called "multicollinearity". Multicollinearity leads to 2 main problems: the coefficients in our fitted model become strongly sensitive to small changes (such as adding or subtracting variables) and the statistical significance of our model also reduces (what makes p-values less reliable). The tool that we use for solving this possible issue with our data is called a correlation matrix. It is simply a symmetric matrix with Pearson correlation coefficients (covariance of 2 corresponding regressors over a product of their standard deviances) in every cell. In our case, these numbers will be represented by colored squares. As we can see in the picture 2, variable ACCESS_CREDIT is strongly correlated with PRODUCT_DEMAND, SHORT_TERM_LIQUIDITY and MEDIUM_TERM_LIQUIDITY. Those last 2 variables are also significantly correlated, what leads us to dismissing both CREDIT_ACCESS aswell as SHORT_TERM_LIQUIDITY. We can begin building and validation of our models with the remaining data .

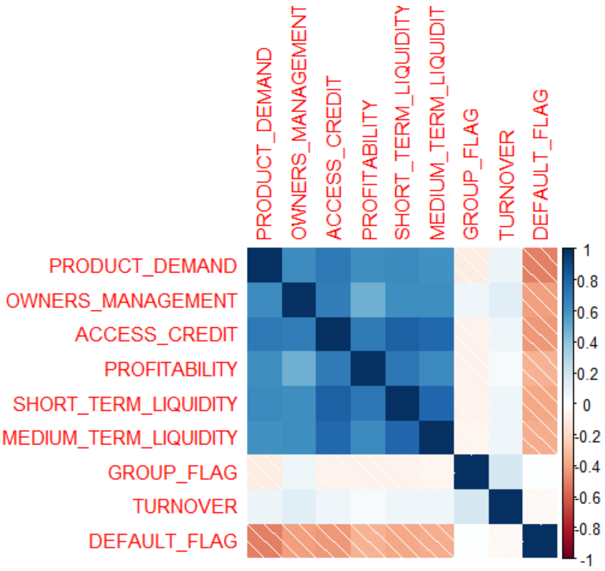


Figure 2: Correlation matrix of regressors

2.2 Data splitting

The important thing to do before proceeding with model fitting is splitting the data into training and test subsets. The training set usually consists of 80% of primary data, and the splitting itself should be executed randomly. The training set will be used to train the models, while the test set will be used to validate them. Such procedure is called "out of sample " forecast. This is also crucial to maintain the distribution proportions of data between these subsets, which can be achieved by using visual methods like histograms or barplots.

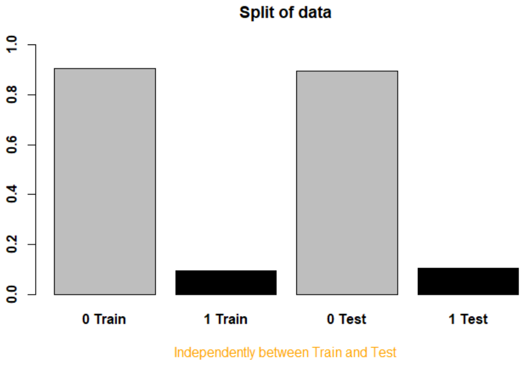


Figure 3: Distribution of DEFAULT_FLAG

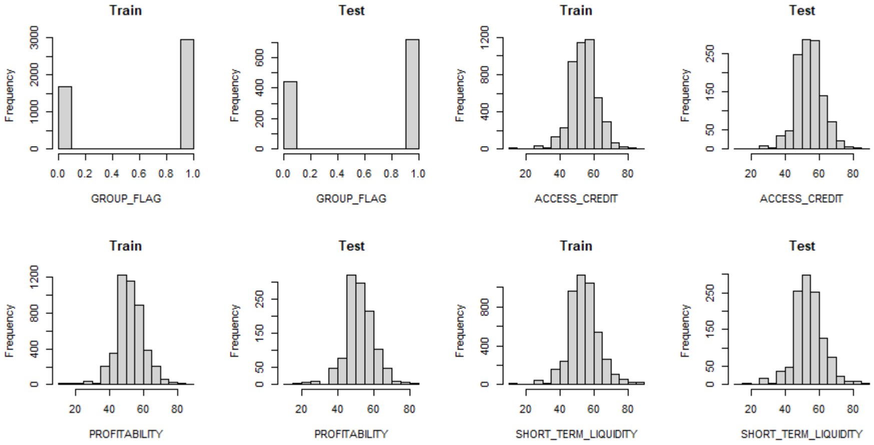


Figure 4: Distribution of some regressors

If we take a look at pictures 3 and 4,
we can see that our variables are equally distributed between the train and test subsets.

3 Logistic regression

Logistic regression is a powerful tool for predicting binary variables like, in our case, DEAFULT_FLAG. It converts numerical values (in a form of linear combinations of regressors) into probabilities of two events (these are numbers from continuous inerval [0,1]). This action is possible thanks to the link function, that "connects" regressors with regressand. For binomial outcomes, there are many different types of link functions. For logistic regression, link function comes in the shape of sigmoid function given in the formula 2:

$$\sigma(X\beta) = \frac{1}{1+e^{-X\beta}}. \tag{2}$$

$X\beta$ in formula 2 is a matrix product of data matrix X - with chosen independant variables as columns, and β - vector of coefficients with values fitted in model building process.

3.1 Full model

Our first model consists of all variables from the basic dataset, including those discarded during data analysis process. We create this model in order to compare it with reduced models. Summary of this model is visible in table 3.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	19.6404	1.0510	18.69	0.0000
PRODUCT_DEMAND	-0.2726	0.0174	-15.69	0.0000
OWNERS_MANAGEMENT	-0.0686	0.0133	-5.16	0.0000
ACCESS_CREDIT	-0.0279	0.0181	-1.54	0.1236
PROFITABILITY	-0.0234	0.0123	-1.91	0.0565
SHORT_TERM_LIQUIDITY	-0.0340	0.0157	-2.17	0.0299
MEDIUM_TERM_LIQUIDITY	0.0048	0.0142	0.34	0.7335
GROUP_FLAG	-0.3000	0.1487	-2.02	0.0436
TURNOVER	-0.0000	0.0000	-0.59	0.5538
INDUSTRYElectricity, Gas and Water	0.6828	0.7442	0.92	0.3589
INDUSTRYExtractive Industries	0.7716	0.7083	1.09	0.2760
INDUSTRYHotels and Leisure	0.3939	0.7116	0.55	0.5800
INDUSTRYManufacturing	0.4922	0.4375	1.13	0.2605
INDUSTRYOffice Machinery and Computer Industries	0.6842	0.4984	1.37	0.1698
INDUSTRYOther	0.5273	0.5216	1.01	0.3120
INDUSTRYProperty and Construction Sectors	1.0125	0.4424	2.29	0.0221
INDUSTRYTrade	0.5558	0.4486	1.24	0.2154
INDUSTRYTransport, Storage and Communications Infrastructure	-0.0916	0.5776	-0.16	0.8740

Table 3: Summary of full logistic model

Column "Estimate" from table 3 is a vector β from formula 2. Its values tell us how many units do the log odds (probability) increase due to the one unit increase of each of these variables.

3.2 Analyst's model

This model is build without use of the variables discarded during data analysis process. Its summary is visible in table 4.

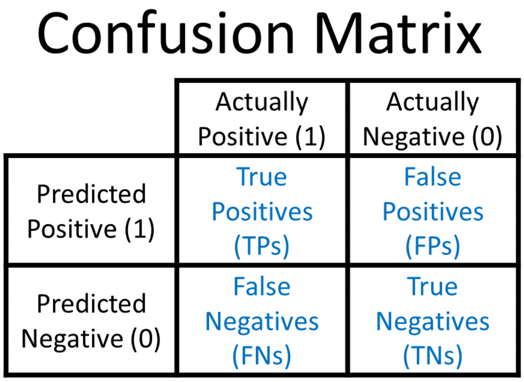
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	19.5792	1.0318	18.98	0.0000
PRODUCT_DEMAND	-0.2754	0.0167	-16.45	0.0000
OWNERS_MANAGEMENT	-0.0860	0.0124	-6.91	0.0000
PROFITABILITY	-0.0396	0.0110	-3.61	0.0003
MEDIUM_TERM_LIQUIDITY	-0.0229	0.0114	-2.01	0.0448
INDUSTRYElectricity, Gas and Water	0.6410	0.7324	0.88	0.3814
INDUSTRYExtractive Industries	0.7582	0.6914	1.10	0.2729
INDUSTRYHotels and Leisure	0.3260	0.7042	0.46	0.6434
INDUSTRYManufacturing	0.4742	0.4324	1.10	0.2728
INDUSTRYOffice Machinery and Computer Industries	0.6632	0.4921	1.35	0.1778
INDUSTRYOther	0.5183	0.5160	1.00	0.3152
INDUSTRYProperty and Construction Sectors	0.9918	0.4368	2.27	0.0232
INDUSTRYTrade	0.5037	0.4435	1.14	0.2561
INDUSTRYTransport, Storage and Communications Infrastructure	-0.0534	0.5660	-0.09	0.9248

Table 4: Summary of Analyst's logistic model

3.3 Automatised model

One attempt to the logistic regression problem is so-called brute-force analysis. The idea is to check all possible subsets of regressors. For obvious reasons, it is not feasible to control each one separately. We conclude that it should be automatised based on some criteria. The first criterion is accuracy, so roughly speaking, the percent of observations that are correctly predicted by the model. To be more precise, we should start with defining confusion matrix. Confusion matrix is a matrix where we distinguish four groups with respect to their actual values and predicted ones. We can see an example of a confusion matrix in figure 5.

As we can see the four groups are TP (True Positive), FN (False Negative), FP (False Positive), TN (True Negative). Columns of the matrix correspond to the actual value of the binary variable, while rows to the predicted one. To calculate the accuracy parameter, we sum correctly predicted values (TP + TN) and divide by a total number of observations (TN + FN + FP + TN). In order to check the performance of the model, the accuracy parameter was calculated on Test subset. The second criterion used by an automatised model evaluation was Akaike information criterion (AIC). The Akaike information criterion is a mathematical method for evaluating how well a model fits the



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 5: Distinguish values with respect to the actual and predicted value

data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data. AIC is calculated from:

- the number of independent variables used to build the model;
- the maximum likelihood estimate of the model (how well the model reproduces the data).

The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables. The exact formula to calculate AIC is given by:

$$AIC = 2K - 2ln(L),$$

where L represents the value of the Maximum Likelihood Function. The automatised program starts with univariate regression. First changes the regressor in next steps takes two regressors then three and so forth until it reaches the full model. The program stores the best model in terms of both accuracy and AIC score. If any of the following models is superior to the stored one, it is stored instead. As a result of the 'brute-force analysis', the final model's summary presents itself in table 10.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	19.8548	0.9292	21.37	0.0000
PRODUCT_DEMAND	-0.2901	0.0164	-17.65	0.0000
OWNERS_MANAGEMENT	-0.0886	0.0124	-7.14	0.0000
MEDIUM_TERM_LIQUIDITY	-0.0373	0.0108	-3.44	0.0006

Table 5: Summary of Automatised logistic model

After a profound analysis, we have concluded that only two levels of' INDUSTRY' may be influential. Due to that fact, we have substituted the variable with dummy variables indicating whether the company is in 'Office Machinery and Computer Industries' or 'Property and Construction Sectors'. The program has chosen the same model, still discarding any indications about 'INDUSTRY' level.

3.4 Cross-Validation

Now we will check how our logistic models perform on the new data (the data that they haven't been trained on). In order to do so, we will use Cross Validation. It is a process in which we randomly divide the training data of our tested model into certain amount of folds (there should be from 5 to 10 folds in order to maintain the balance between bias and variance of our models). Then we fit the model on union of all folds except one. We repeat this operation for every fold and we receive accuracy and kappa statistic from every repetition. The kappa statistic is a measurement of the accuracy of a model, while taking into account chance (the closer the value is to 1 the better). Eventually, we take these statistics and calculate mean value from them. In our analysis, we perform 10-fold repeated cross validation, which means that we divide the training set into 10 folds and after finishing one validation, we perform another on different 10 folds of the training set. The results are visible in the table 6.

Model name	Accuracy	Kappa
Full model	0.926	0.460
Analyst's model	0.925	0.448
Automatised model	0.924	0.440

Table 6: CV results for logistic regression

We can observe that the models get very similar results. Even though a slight advantage suggests to use Analyst's model, we may hold to using Automatised model, because it uses less regressors. However one may still debate which one is superior.

3.5 Probit regression

Another useful tool, when it comes to the prediction of binary variables is probit regression. It is very similar to the logistic regression approach, as it also provides us with probabilities of success (in terms of predicted variable) for certain observations (rows of data matrix X from formula 2. The main difference is the link function, which in this case is a Cumulative Distribution Function of the standard normal distribution. In other words, we can treat $X\beta$ from formula 2 as a quantile of the standard normal distribution. The probability of success $P(Y = 1|X)$ (where Y is our regressand) is then evaluated with formula 3:

$$P(Y = 1|X) = \Phi(X\beta) \tag{3}$$

Usually, using probit or logistic regression does not make a big difference, but probit is believed to behave more accurately when dealing with extremaly independent regressors, which in our case may be crucial, as our variables are more or less correlated with each other. Another difference is that in probit regression $\Phi(X\beta)$ values approach borders (0 or 1) a bit faster. For our analysis, we build model with the use of variables from Automatised model. Its summary is visible in table 7.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	8.6617	0.4106	21.09	0.0000
PRODUCT_DEMAND	-0.1375	0.0078	-17.67	0.0000
OWNERS_MANAGEMENT	-0.0360	0.0062	-5.78	0.0000
MEDIUM_TERM_LIQUIDITY	-0.0132	0.0055	-2.39	0.0167

Table 7: Summary of probit model

4 Linear regression

IT department has raised concerns about the usability of a logistic regression model as it represents some issues from a systems implementation standpoint. In this section, we present the results of linear modelling analysis. From a mathematical point of view, logistic regression is much more reasonable while predicting binary regressand than the linear model. However, the issue will be addressed more profoundly in the model comparison section. As for now, we would like only to present results of the linear modelling. Similarly to logistic modeling, we decided to create three models with different approaches. First, the most basic one is a full model. The results are presented in table 8. It is important to note that linear models often predicted a negative probability. To make rational inferences negative probabilities are substituted with zeros. The issue concerns all presented linear models. It is particularly important when calculating AUC (refer to section 5.1).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3097	0.0378	34.66	0.0000
PRODUCT_DEMAND	-0.0142	0.0008	-18.62	0.0000
OWNERS_MANAGEMENT	-0.0039	0.0007	-5.68	0.0000
ACCESS_CREDIT	-0.0025	0.0009	-2.65	0.0080
PROFITABILITY	-0.0002	0.0006	-0.23	0.8151
SHORT_TERM_LIQUIDITY	-0.0017	0.0008	-2.25	0.0246
MEDIUM_TERM_LIQUIDITY	0.0009	0.0007	1.23	0.2189
GROUP_FLAG	-0.0270	0.0079	-3.40	0.0007
TURNOVER	0.0000	0.0000	1.82	0.0681
INDUSTRYElectricity, Gas and Water	0.0624	0.0379	1.65	0.1000
INDUSTRYExtractive Industries	0.0559	0.0386	1.45	0.1478
INDUSTRYHotels and Leisure	0.0252	0.0384	0.66	0.5109
INDUSTRYManufacturing	0.0233	0.0212	1.10	0.2721
INDUSTRYOffice Machinery and Computer Industries	0.0460	0.0251	1.83	0.0675
INDUSTRYOther	0.0419	0.0267	1.57	0.1165
INDUSTRYProperty and Construction Sectors	0.0801	0.0222	3.61	0.0003
INDUSTRYTrade	0.0241	0.0219	1.10	0.2709
INDUSTRYTransport, Storage and Communications Infrastructure	-0.0028	0.0266	-0.10	0.9175

Table 8: Linear regression full model

4.1 Automatised model

Our second attempt to create the best linear model is 'brute-force' analysis. Similarily to logistic regression, the program checks all possible subsets of variables to consider as regressors. Based on accuracy score and AIC the program estimates the best model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2110	0.0362	33.49	0.0000
PRODUCT_DEMAND	-0.0196	0.0005	-38.58	0.0000
GROUP_FLAG	-0.0339	0.0078	-4.35	0.0000
INDUSTRYElectricity, Gas and Water	0.0607	0.0383	1.59	0.1126
INDUSTRYExtractive Industries	0.0611	0.0390	1.57	0.1169
INDUSTRYHotels and Leisure	0.0294	0.0387	0.76	0.4482
INDUSTRYManufacturing	0.0249	0.0214	1.16	0.2444
INDUSTRYOffice Machinery and Computer Industries	0.0518	0.0253	2.05	0.0409
INDUSTRYOther	0.0460	0.0269	1.71	0.0870
INDUSTRYProperty and Construction Sectors	0.0864	0.0223	3.87	0.0001
INDUSTRYTrade	0.0250	0.0221	1.13	0.2575
INDUSTRYTransport, Storage and Communications Infrastructure	0.0059	0.0268	0.22	0.8268

Table 9: Summary of Automatised linear model

4.2 Analyst's model

It was clear that many of the variables are unserviceable. Theirs t-value suggested total insignificance in the model. Also, again it seemed unambiguous that only two levels of 'INDUSTRY' factor may have been significant. To avoid linear correlation, 'INDUSTRY' variable was substituted with two specific levels of the factor, that is 'Office Machinery and Computer Industries' and 'Property and Construction Sectors'. It became clear that our attempt with 'brute-force' analysis was impeded by 'INDUSTRY' factor. The program was not allowed to adjust variable in a sense that to specify a few levels. The results of another run of the program on the new data look much better.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2365	0.0307	40.25	0.0000
PRODUCT_DEMAND	-0.0196	0.0005	-38.59	0.0000
GROUP_FLAG	-0.0331	0.0077	-4.27	0.0000
INDUSTRY_Property_and_Construction_Sectors	0.0591	0.0097	6.12	0.0000

Table 10: Summary of automatised linear model on adjusted data

The model is taking only three regressors. We may already say that it is not overfitted. We have taken one more attempt at the model. This time we added a few variables. To avoid a high correlation between variables 'ACCESS_CREDIT' was discarded, although it seemed to be heavily significant. 'OWNERS_MANAGEMENT' and 'SHORT_TERM_LIQUIDITY' were added to the model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3349	0.0323	41.38	0.0000
PRODUCT_DEMAND	-0.0147	0.0007	-20.95	0.0000
OWNERS_MANAGEMENT	-0.0042	0.0006	-6.62	0.0000
SHORT_TERM_LIQUIDITY	-0.0024	0.0005	-4.49	0.0000
GROUP_FLAG	-0.0230	0.0078	-2.96	0.0031
INDUSTRY_Property_ and_Construction_Sectors	0.0556	0.0096	5.80	0.0000

Table 11: Summary of Analyst's linear model

4.3 Experts' model

Several credit experts came up with the following model,

$$PD = \frac{1}{1 + e^{-0.1*\text{Score}}}$$

where the Score is calculated as

$$\begin{aligned}\text{Score} = 0.2 * \text{'Product and Demand'} + 0.1 * \text{'Quality of management'} + \\ +0.1 * \text{'Access to Credit'} + 0.15 * \text{'Profitability'} + \\ +0.25 * \text{'Ability to pay'} + 0.2 * \text{'Solvency'}\end{aligned}$$

The model's performance is unacceptable. The lowest obtained PD is 0.731. Assuming the threshold level on 0.5 all applicants are forecasted to default. Score using AUC (refer to section 5.1) is barely 0.11. We decide not to perform further analysis of the model and discard it on the spot.

5 Model comparison

5.1 ROC and AUC

Receiver operating characteristic (ROC) is is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity. The false-positive rate is also known as the probability of false alarm. Referring to symbols introduced in section 3.3, TPR is given by the formula TPR $= \frac{TP}{N} = \frac{TP}{TP+FN}$. FPR can be calculated as FPR $= \frac{FP}{N} = \frac{FP}{FP+TN}$.
In our case, threshold varied from zero to one by 0.005. The last number was chosen in order to average between accuracy and computing time. The prediction was assigned by comparison predicted PD with a given threshold.

Area under curve (AUC) is calculated concerning the trapezoidal rule. It is the numerical interpretation of ROC.
ROC and calculated AUC calculated for particular models are presented in detail in Appendix B (section 8).

5.2 Model comparison

In the final stage of our project, we compare all of our previously built models with the use of already introduced tools - AUC and Accuracy. The process of validation is conducted on the Test set. All results are presented in table 12. Having analysed it, we can tell that logistic models are a bit more accurate than the linear ones, which is not quite surprising, as we are dealing with binary regressand. It is also worth noticing, that models with all variables included (both logistic and linear) have the biggest AUC of all models, but they have lower accuracies compared to their modifications. They can be interpreted to be the best at predicting most of the true defaults, at the cost of greater probability of false alarm. There is also a substantial increase in AUC for the linear models after data manipulation. It is the result of replacing factor INDUSTRY with dummy variables of its two most statistically significant levels "Office Machinery and Computer Industries" and "Property and Construction Sectors". However, the model lost some Accuracy. The same scenario concerns the probit model, but in this case, model lost a lot of Accuracy in exchange for small increase in AUC. That makes the Automatised logistic model the most accurate in our whole analysis, and thus we choose this one as the best for predicting Probability of Default.

Model name	Accuracy	AUC
Full logistic model	0.9156	0.9687
Analyst's logistic model	0.9182	0.9667
Automatised logistic model	0.9199	0.9643
Automatised probit model	0.9104	0.9659
Full linear model	0.8992	0.9682
Analyst's linear model	0.9001	0.9667
Automatised linear model	0.9018	0.9616
Expert's model	0.1059	0.1117

Table 12: Overall Validation Comparison

The Head of the Credit team has requested to conduct a segmentation analysis by Financial Holding Type. To do so, we create two logistic and two linear models for both holding and subsidiary companies separately. Their summaries are visible in tables 13-16.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	23.3426	1.8923	12.34	0.0000
PRODUCT_DEMAND	-0.3614	0.0368	-9.81	0.0000
ACCESS_CREDIT	-0.1287	0.0234	-5.51	0.0000
INDUSTRY_Property_and_ Construction_Sectors	0.7498	0.2847	2.63	0.0084

Table 13: Summary of logistic model for holding companies

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	17.7685	1.0404	17.08	0.0000
PRODUCT_DEMAND	-0.2544	0.0195	-13.06	0.0000
OWNERS_MANAGEMENT	-0.0727	0.0148	-4.93	0.0000
ACCESS_CREDIT	-0.0030	0.0201	-0.15	0.8825
SHORT_TERM_LIQUIDITY	-0.0480	0.0167	-2.87	0.0041

Table 14: Summary of logistic model for subsidiary companies

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2902	0.0490	26.36	0.0000
PRODUCT_DEMAND	-0.0205	0.0008	-25.10	0.0000
INDUSTRY_Property_and_ Construction_Sectors	0.0755	0.0163	4.63	0.0000

Table 15: Summary of linear model for holding companies

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2938	0.0401	32.25	0.0000
PRODUCT_DEMAND	-0.0140	0.0009	-15.04	0.0000
OWNERS_MANAGEMENT	-0.0047	0.0008	-5.59	0.0000
ACCESS_CREDIT	-0.0024	0.0009	-2.62	0.0088
INDUSTRY_Property_and_ Construction_Sectors	0.0491	0.0119	4.13	0.0000

Table 16: Summary of linear model for subsidiary companies

Model name	Accuracy	AUC
Logistic for holding companies	0.9209	0.9634
Logistic for subsidiary companies	0.921	0.9667
Linear for holding companies	0.9074	0.9595
Linear for subsidiary companies	0.8997	0.9699

Table 17: Comparison of models by financial holding type

After analysing the performance from table 17, we can observe that both linear and logistic models gained on Accuracy, while preserving the same level of AUC (compared to their counterparts from table 12). It may seem, that using these two bespoke models (especially logistic ones) might be worth considering from qualitative point of view.

6 References

Bibliography

[1] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.329.4866rep=rep1type=pdf

[2] https://stats.stackexchange.com/questions/20523/difference-between-logit-and-probit-models

[3] https://www.scribbr.com/statistics/akaike-information-criterion/

[4] https://glassboxmedicine.files.wordpress.com/2019/02/confusion-matrix.png

[5] https://en.wikipedia.org/wiki/Receiver_operating_characteristic

[6] HSBC Quant Academy lectures and .R files

7 Appendix A

Appendix A presents graphical Information Value for variables with respect to regressand 'DEFAULT_FLAG'.

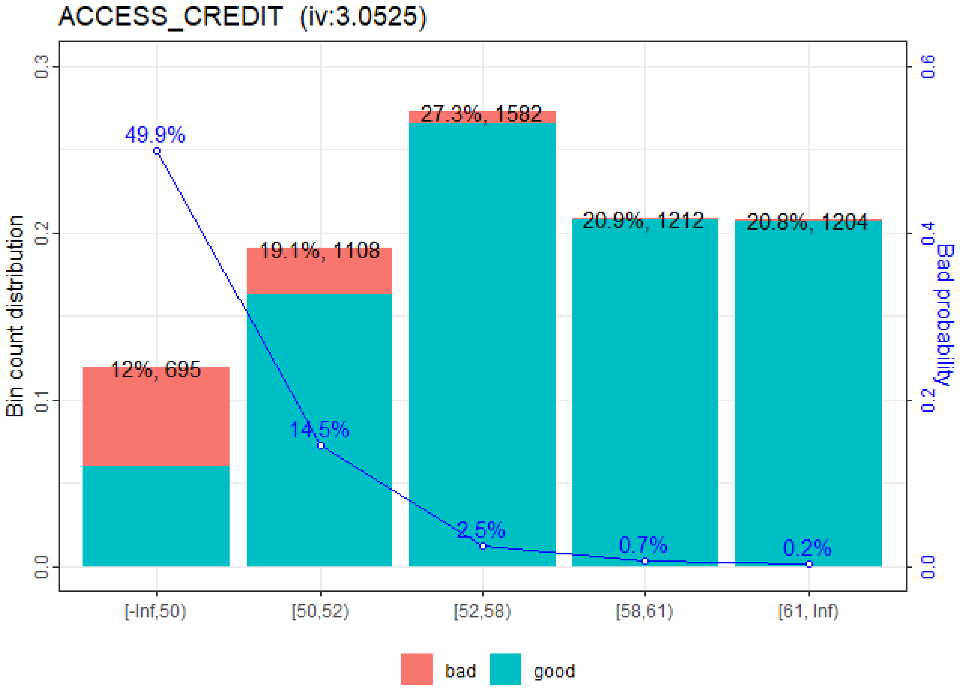


Figure 6: IV Access Credit

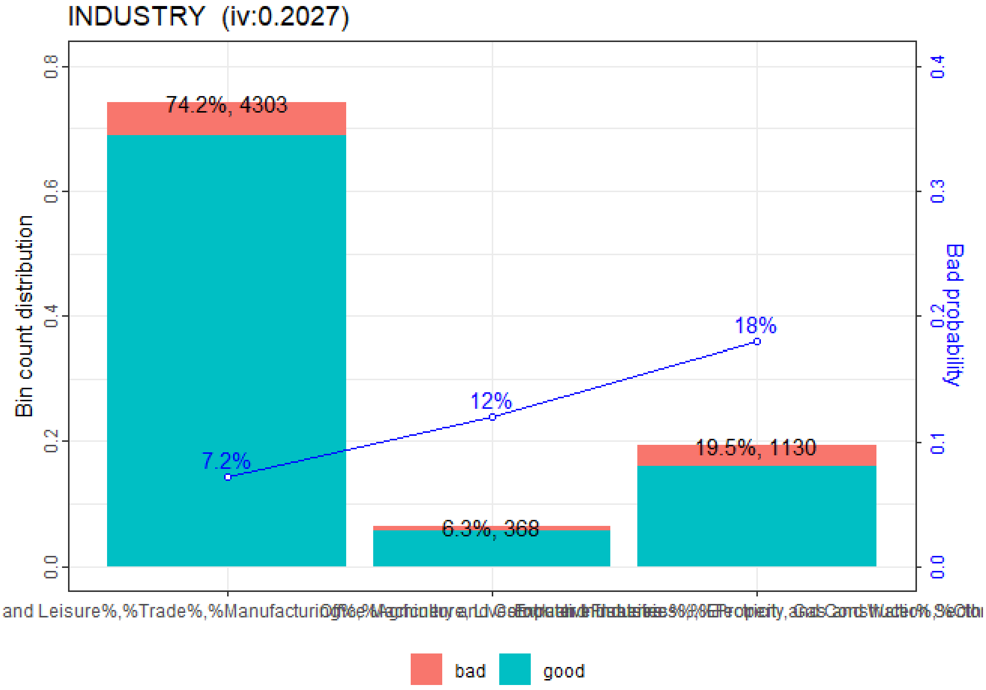


Figure 7: IV Industry

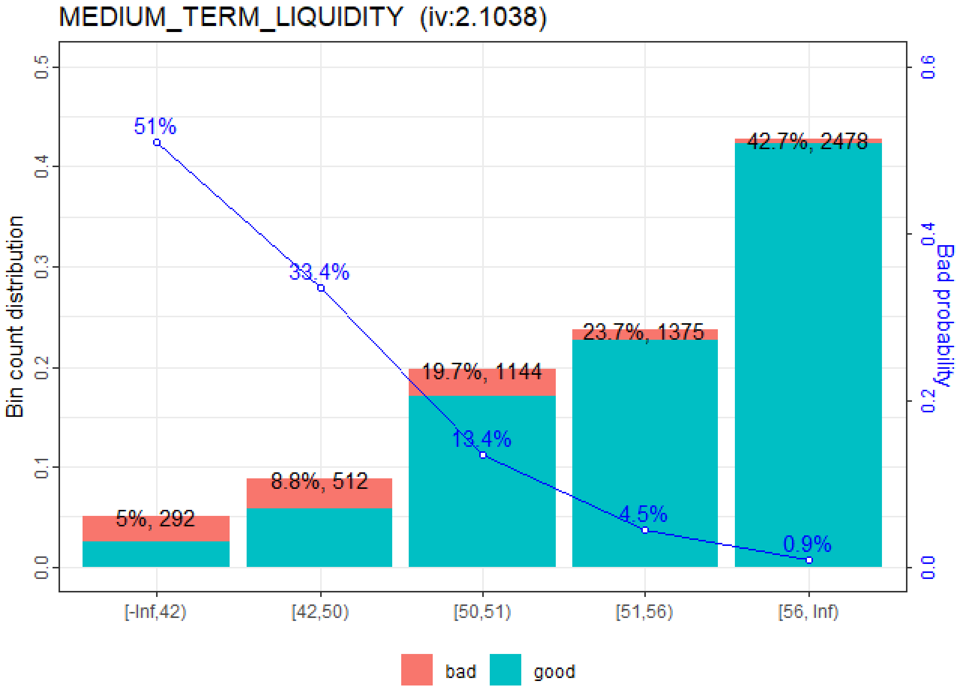


Figure 8: IV Medium Term Liquidity

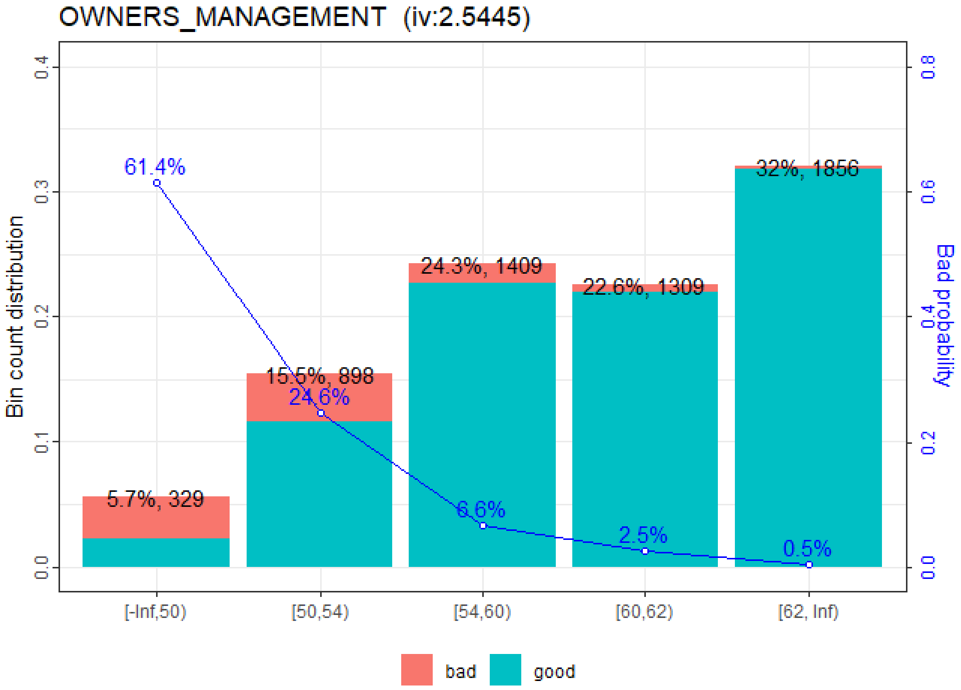


Figure 9: IV Owners Management

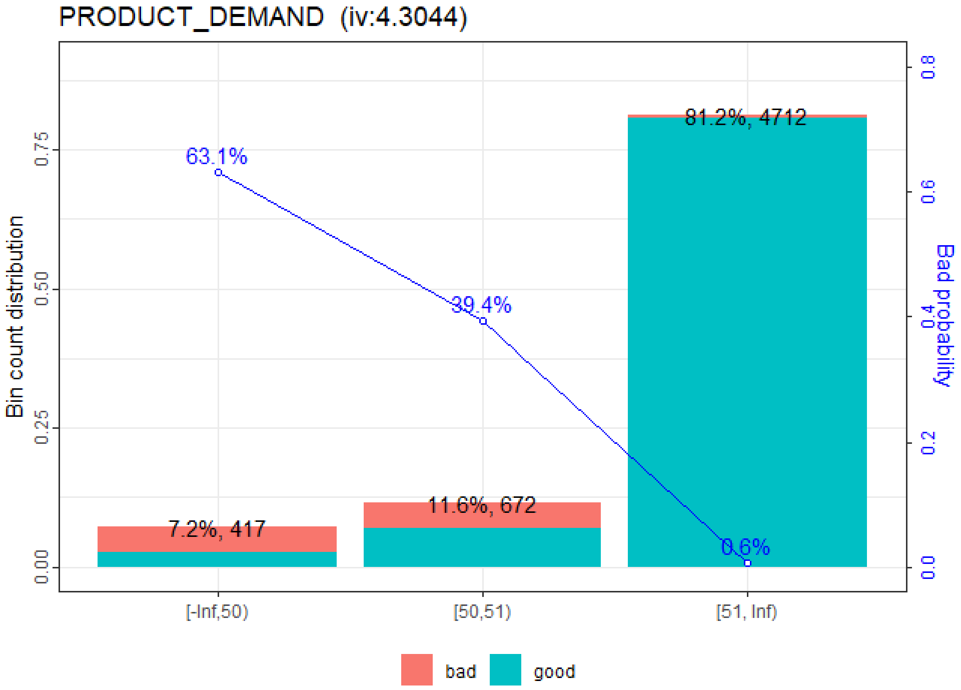


Figure 10: IV Product Demand

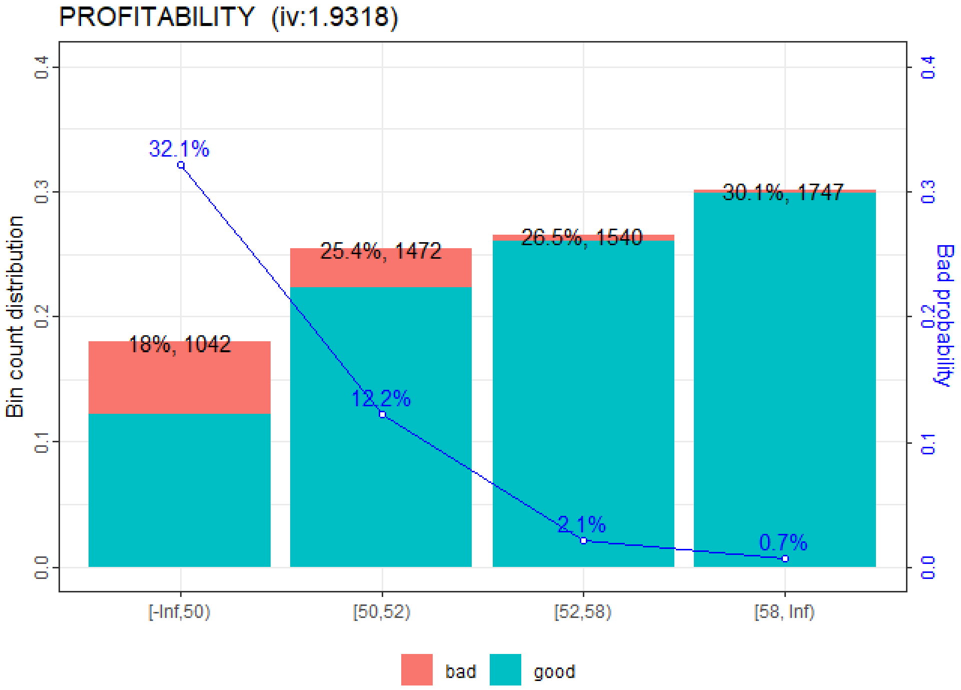


Figure 11: IV Profitability

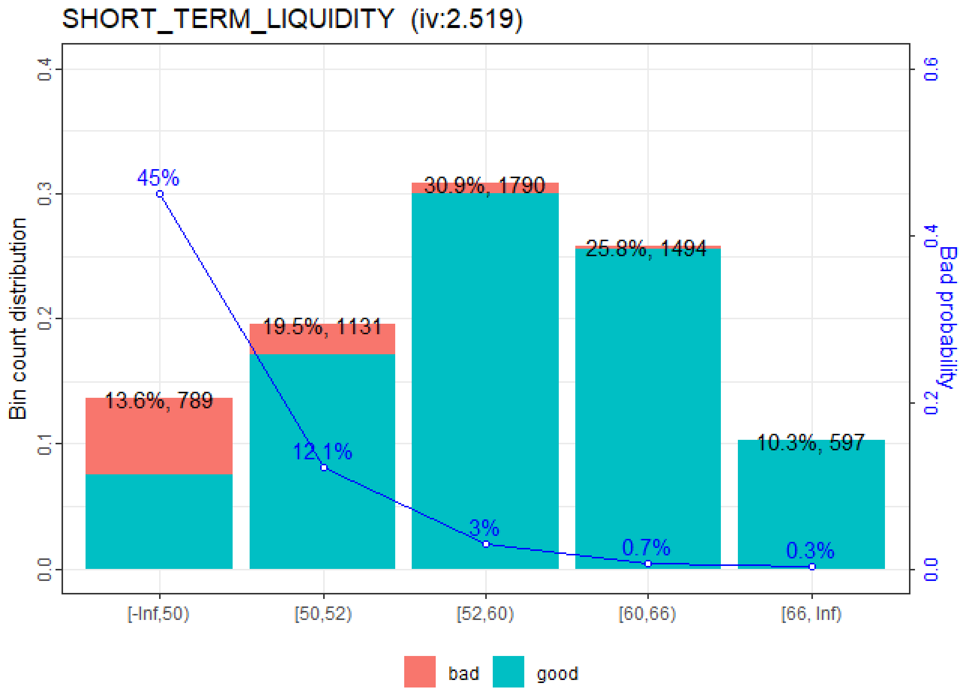


Figure 12: IV Short Term Liquidity

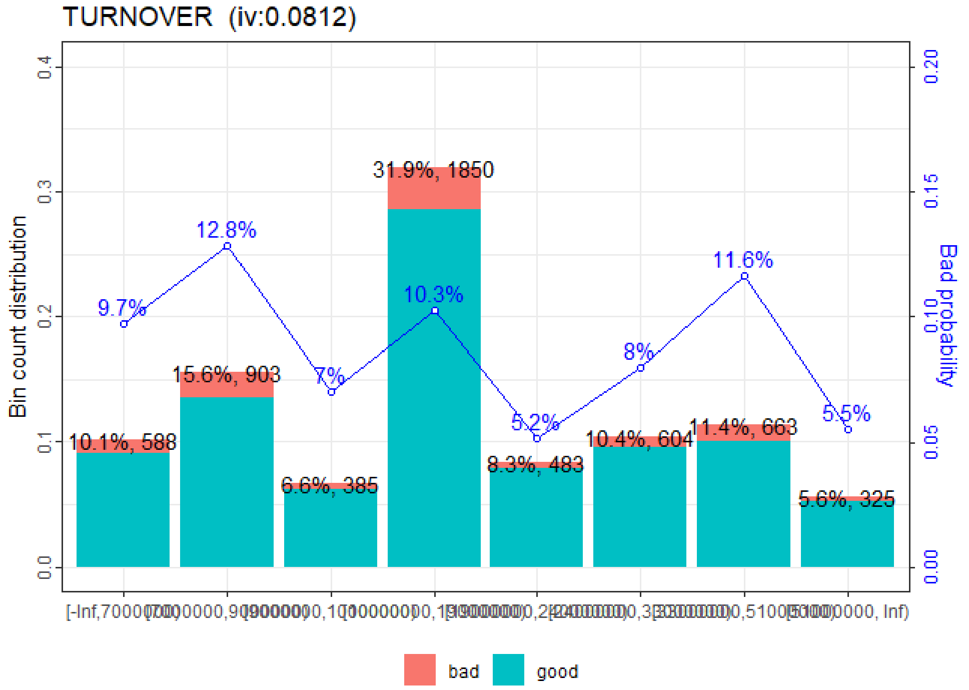


Figure 13: IV Turnober

8 Appendix B

In the appendix, we present ROCs with calculated AUC for particular models.

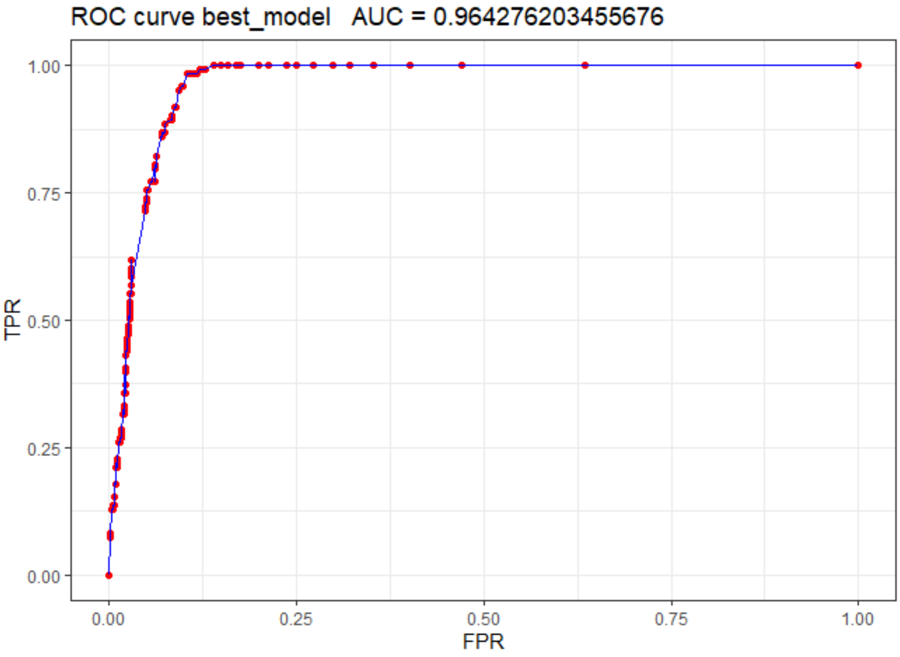


Figure 14: Automatised logistic model

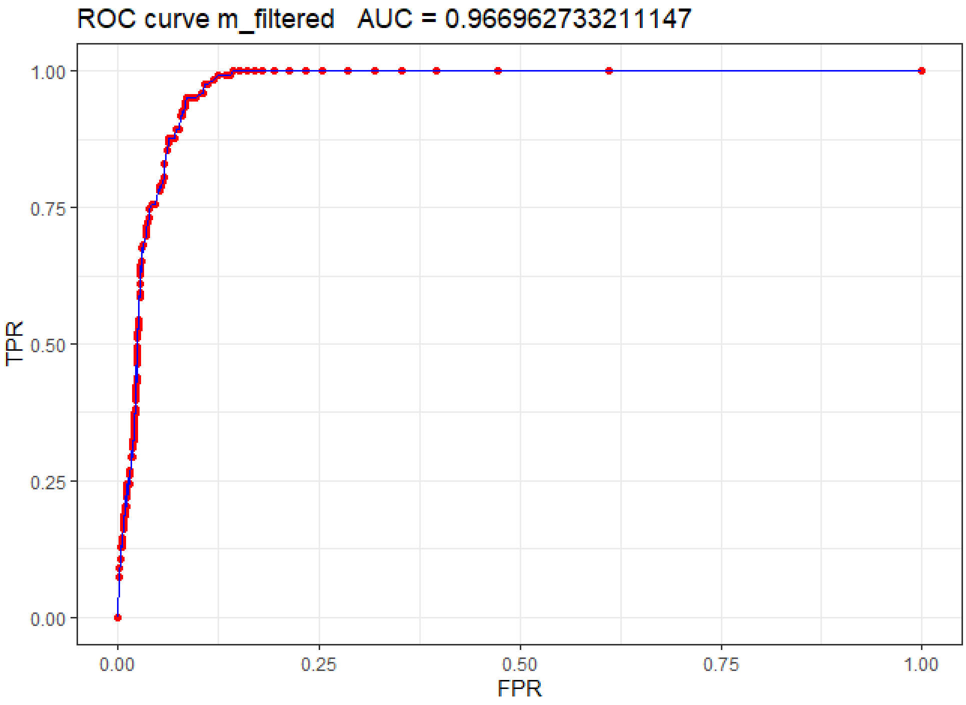


Figure 15: Analyst's logistic model

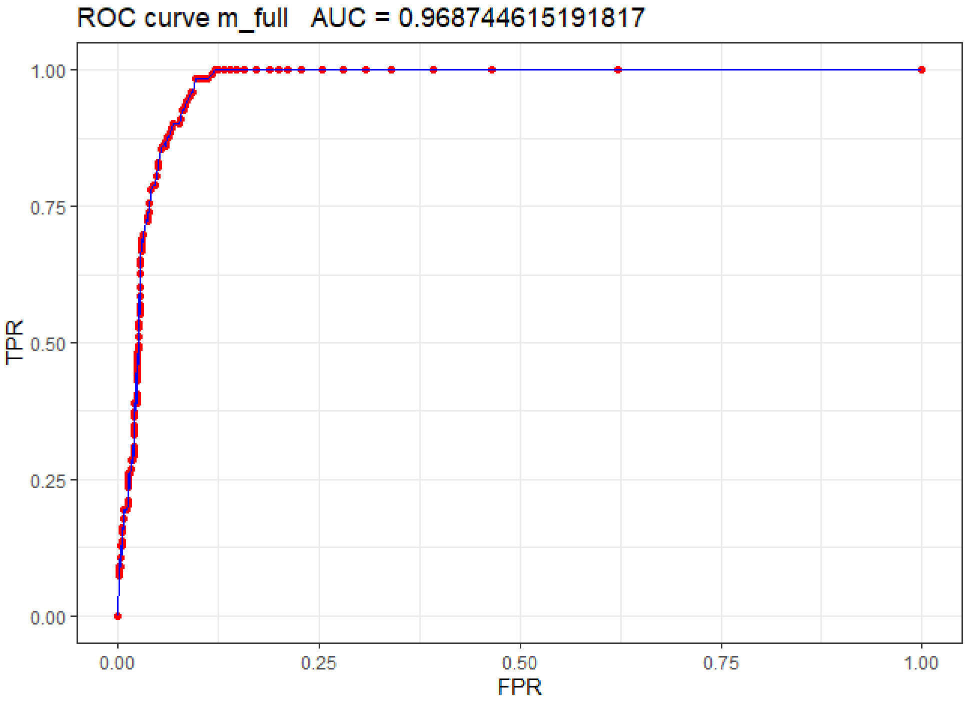


Figure 16: Full logistic model

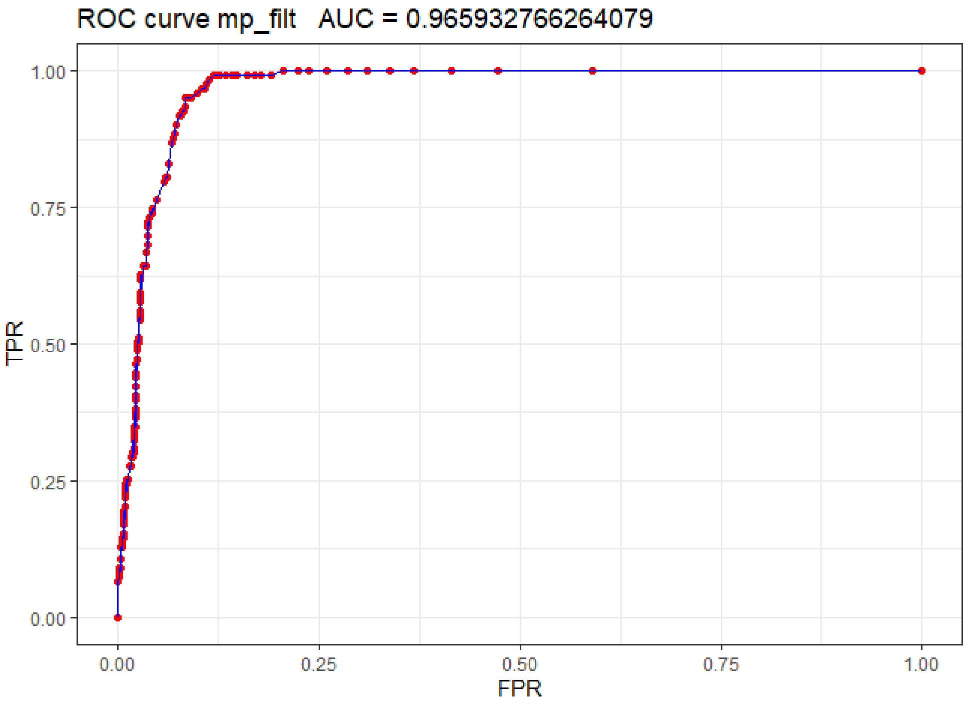


Figure 17: Analyst's probit model

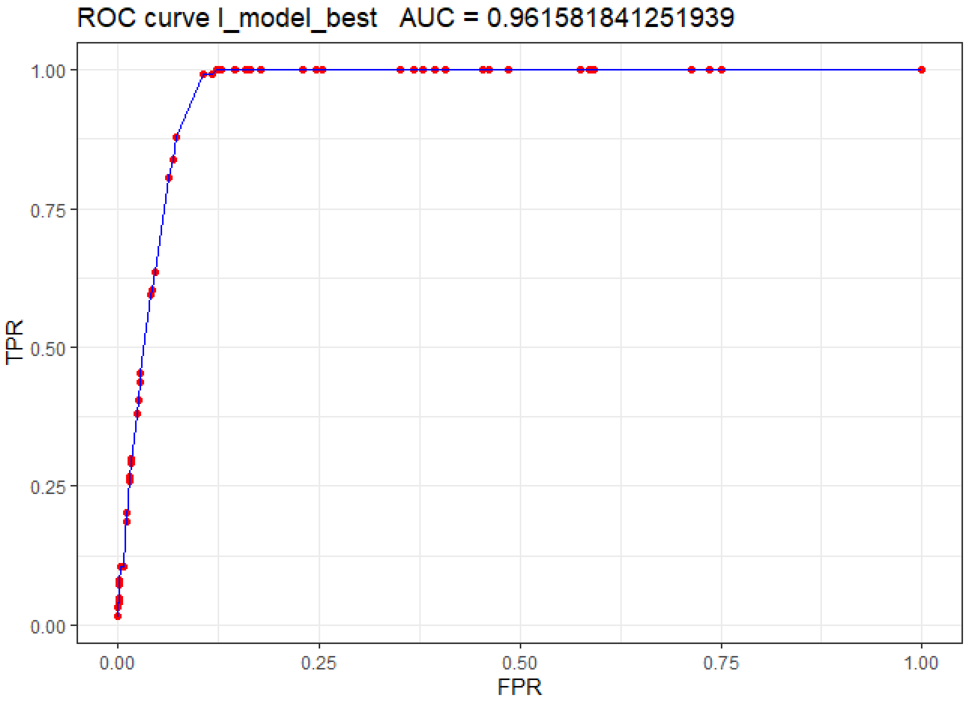


Figure 18: Automatised linear model

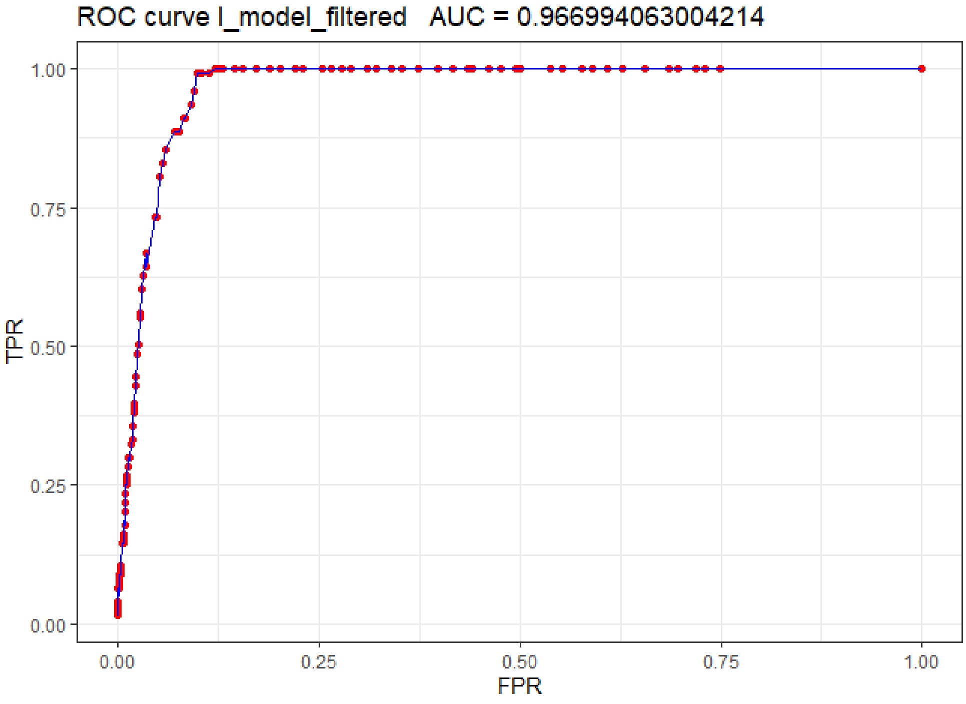


Figure 19: Analyst's linear model

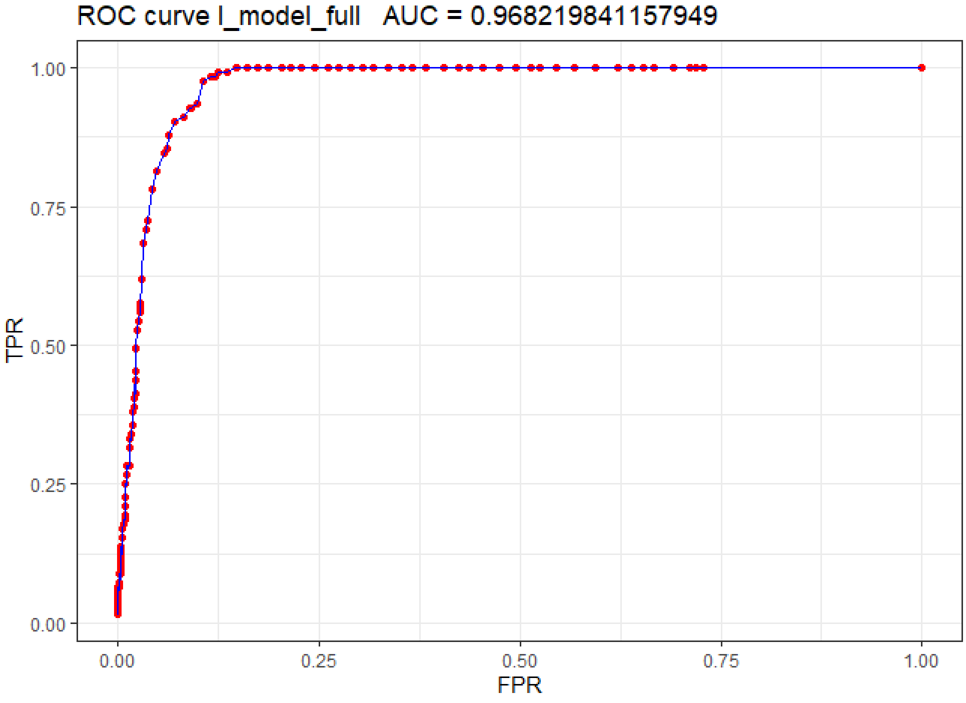


Figure 20: Full linear model

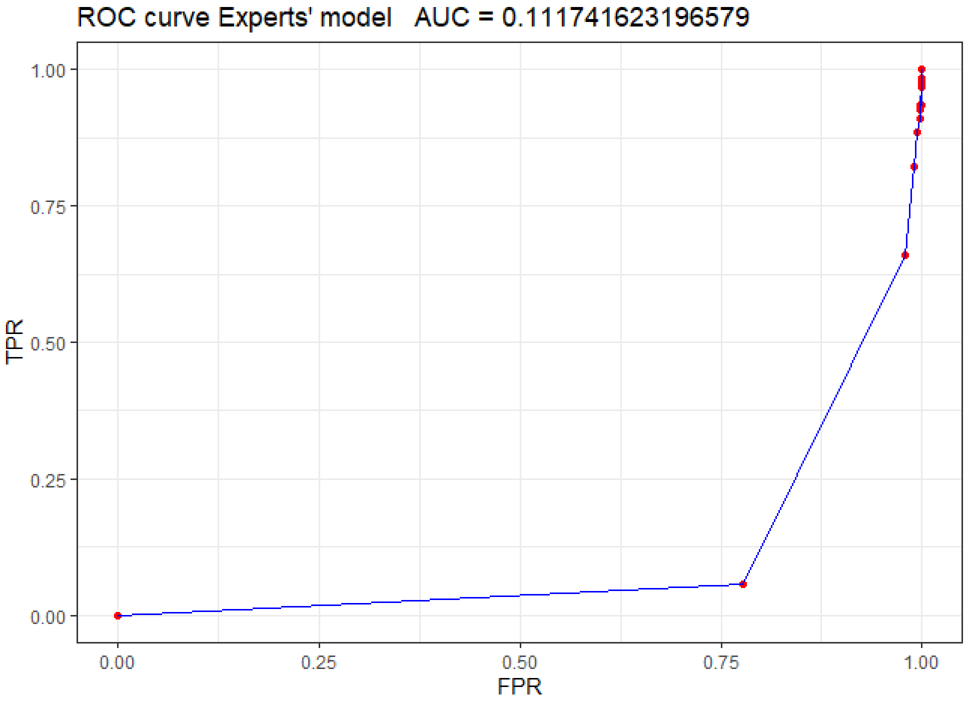


Figure 21: Experts' model